EPA Comments on LWG Benthic Approach

EPA combined comments on Section 6 and Attachments 6 and 7 of the Draft Baseline Ecological Risk Assessment, Benthic Approach Reanalysis Technical Memorandum and Site Specific SQGs based on Individual Bioassay Endpoints

Risk Management

1) The BERA is not a competition between multiple lines of evidence or approaches of estimating sediment toxicity to benthic biota, with the winner being the most reliable. Although the primary lines of evidence for evaluating benthic risk are the empirical sediment bioassays, each of the three primary lines of evidence evaluating benthic risk from sediment chemistry (bulk sediment chemistry guidelines such as PECs, logistic regression models and floating percentile models) provide different information to EPA risk assessors and risk managers. Sediment quality guidelines provide information about sediment chemical concentrations associated with adverse effects to benthic biota at other locations in North America. Logistic regression models provide information about the probability of toxicity to benthic biota from mixtures of chemicals. The floating percentile model provides information about predicted error rates of sediment benchmarks derived from organisms exposed to mixtures of chemicals. Findings from all three of these lines of evidence must be carried through to the completion of the risk characterization phase of the BERA. The reliability analysis can be used in the uncertainty analysis to assist the development of chemistry based cleanup levels protective of the benthic community in the feasibility study.

2) The risk characterization conclusions, including all hazard quotient calculation results and their uncertainties from all three lines of evidence (generic sediment quality benchmarks, floating percentile model, logistic regression model) must be reported in the final BERA. Use of reliability statistics in the BERA to eliminate lines of evidence or individual sediment quality benchmarks from risk analysis and risk characterization is unacceptable to EPA.

3) EPA expects description of predictive model and sediment quality benchmark uncertainties to be the primary use of reliability statistics in the BERA. Identification of models and benchmarks that maximize the agreement between predicted and measured toxicity (i.e. simultaneously minimize both false positives and false negatives) is also a valid use of reliability statistics in the BERA. Use of reliability statistics to eliminate individual chemical benchmarks or entire lines of evidence is not acceptable to EPA in the BERA, although the reliability information can and likely will be used by EPA risk managers during their remedy selection process.

4) The risk analysis appears to rely on inappropriate application of tools and/or inappropriate interpretation of the results. Such decisions on risk assessment methodology minimize the size of the areas that pose unacceptable risks to the benthic community, rather than the objective presentation of areas posing potentially unacceptable risks that is required to be presented in the BERA.

5) Although EPA does not agree with the conclusions regarding the proportion of the site area with potentially unacceptable benthic risks (we believe the area with potentially unacceptable

risks are underestimated), the approach used by LWG to determine the spatial extent of these areas is sound. It will provide useful information to both EPA risk assessors and risk managers once the empirical sediment toxicity data is fully incorporated to the approach using the correct definitions of the four different levels of toxicity as defined by the site specific reference envelope approach, and the final derivation of site specific, risk based sediment quality benchmarks. EPA believes that chemicals in addition to PCBs, PAHs and the DDT/DDD/DDE potentially pose unacceptable risks to benthic biota, these three chemicals are not the only COCs potentially posing unacceptable risks. Once the benthic toxicity reanalysis has been fully completed, the full list of COCs must be presented in the BERA.

6) The draft BERA states "None of the generic SQGs could reliably predict toxicity in Portland Harbor sediments (Attachment 7); therefore, the generic SQGs were not used in risk characterization for the BERA." This conclusion is incorrect and should be removed from the revised BERA. USEPA expects that sediment chemistry data, as evaluated using the generic SQGs, will be included in the risk characterization for benthic invertebrates. Published reliability criteria for generic SQGs such as probable effect concentrations (PECs, MacDonald et al. 2000) largely meet LWG's proposed reliability criteria.

7) The elimination of the generic SQG line of evidence for benthic invertebrates is unwarranted. The reliability of predicted benthic toxicity using the Portland Harbor dataset may be influenced by many factors, including the methods used to evaluate toxicity in site-specific samples. Consistent with EPA's February 15, 2008 Problem Formulation, the SQG line of evidence must be presented in the draft BERA. The predictability evaluation should be discussed in the uncertainty analysis.

8) The interpretation of benthic toxicity results as described in the Benthic Toxicity Reanalysis Technical Memorandum and Attachment 6 of the draft BERA are unacceptable. Errors include not using only the Portland Harbor specific reference envelope, not describing and using the four groups of toxicity consistently (Levels 0, 1, 2, and 3), and the use of reliability statistics to eliminate generic sediment quality guidelines and the logistic regression results from the risk characterization sections of the BERA. All sediment lines of evidence must be fully evaluated, and none can be eliminated from the BERA due to perceived lack of reliability issues. EPA's detailed comments on reliability analyses, provided separately, provides the reliability measures EPA expects LWG to use in the next draft of the BERA.

Bioassays

1) The interpretation of the sediment toxicity data in the BERA is incorrect, as only the Level 3 responses were identified as toxic. Sediment samples with Level 1 and Level 2 response rates should also be designated as toxic in the BERA, in accordance with USEPA guidance. Level 1, Level 2 and Level 3 response rates should be designated as low, moderate and high toxicity respectively as specified in MacDonald Landurm (2008). In addition, the Benthic Toxicity Reanalysis Tech Memo describes three procedures for developing the reference envelope to classify sediment samples as toxic and not toxic. Two procedures (Calcasieu and Draft BERA) are inconsistent with previous EPA comments and direction and should be eliminated. The

Evaluation of sediment toxicity in the revised BERA should be based solely on the EPA 2009 reference envelope approach.

2) Within the BERA, the areas posing potentially unacceptable risks to benthic biota are the areas with Level 1, Level 2 or Level 3 toxicity to any one of the four sediment toxicity test endpoints, not just areas with Level 3 toxicity. As a result, the spatial evaluation of the proportion of the site with potentially unacceptable risks is expected to be greater than 5%,. The specific areas identified in the text as posing potentially unacceptable risks to benthic biota are correctly identified as among the areas with the highest risks, however, several other areas of potentially unacceptable risks, such as the mouth of Swan Island Lagoon (RM 8.2, east bank), have also been identified by EPA as having elevated ecological risks to benthic biota. A series of maps depicting areas of sediment toxicity based on an interpolation of bioassay results is attached. Consistent with previous EPA comments and direction, the risk assessment should identify Level 1 (low), Level 2 (moderate) and Level 3 (high) toxicity as posing potentially unacceptable risk. EPA acknowledges our previous determination that Level 2 and Level 3 risks constitute unacceptable risks for the purposes of the Portland Harbor FS.

3) The risk questions that were developed to link the measurement endpoints to the assessment endpoint for benthic invertebrates are, in some cases, incorrect. For example, the first risk question should read, “Is the survival or biomass of benthic invertebrates, as evaluated using the results of whole sediment toxicity tests with the amphipod, *Hyalella azteca,* and the midge, *Chironomus dilutus*, below the lower limit of the reference envelope?” The second risk question should read, “Do the concentrations of COPCs in surficial sediments exceed the generic and/or site-specific toxicity thresholds that are selected for classifying sediment samples as likely toxic or likely not toxic to benthic invertebrates, considering the survival, growth, and/or reproduction endpoints?”

4) The discussion of the rationale for using total biomass as the measurement endpoint in some sediment toxicity tests needs to be amended to remove the impression that the total biomass measure is a non-standard endpoint. More recent experience with sediment toxicity tests within North America in recent years has indicated that total biomass reductions often serve to eliminate a common problem with the growth endpoint, which is that the remaining biomass in individual organisms where some mortality has occurred is higher than the individual biomass in control or reference organisms, indicating enhanced growth due to less competition for food resources because fewer individuals are in the test sediments. The actual interpretation of such results is that some level of toxicity has occurred due to a reduction in the test population's total biomass. Table 6-2 of the BERA needs to be replaced with the correct reference envelope definitions for the BERA previously provided to LWG by EPA. In addition, Table 6-2 should include four categories of toxicity, not the three as shown. The four correct categories of toxicity based on the reference envelope approach are Level 0 (no toxicity), Level 1 (low toxicity), Level 2 (moderate toxicity), and Level 3 (high toxicity). Although EPA reviewers understand the Level 0, 1, 2 and 3 terminology, readers not versed in the details of sediment toxicity testing, including the general public, do not understand these definitions, thus the need for the no, low, moderate and high definitions. Footnote C of Table 6-2 should be corrected to show the endpoint is total biomass, not growth. The appropriate sections of Attachment 6 also need corrected.

5) The counts of the empirical bioassay results that fall into each of the four classes of toxicity (Levels 0, 1, 2 and 3) do not match the counts identified by EPA. Table 6.3 should be corrected to reflect the prevalence of toxicity as follows (based on the 293 stations used to develop predictive toxicity models). **Chironomus survival:** Level 0 = 188 stations; Level 1 = 54 stations; Level 2 = 19 stations; Level 3 = 32 stations. **Chironomus biomass:** Level 0 = 201 stations; Level 1 = 37 stations; Level 2 = 12 stations; Level 3 = 43 stations. **Hyalella survival:** Level 0 = 253 stations; Level 1 = 19 stations; Level 2 = 2 stations; Level 3 = 19 stations. **Hyalella biomass:** Level 0 = 167 stations; Level 1 = 53 stations; Level 2 = 43 stations; Level 3 = 30 stations.

6) The draft BERA seems to imply that the reference envelope approach used in Portland Harbor is a non-standard approach. However, the discussion is based on older literature, and does not incorporate knowledge gained at multiple sites in recent years with the benefits of the reference envelope approach as applied to Portland Harbor. More recent experience with reference envelopes by some of the leading experts in the field of sediment toxicity testing is the basis for the Portland Harbor approach, as it provides benefits over earlier reference envelope approaches. The text needs to be amended to reflect this newer knowledge.

7) The Technical Memorandum concludes that the *Hyalella* biomass endpoint should be interpreted with caution. While challenges with the negative control treatments are correctly identified in the Benthic Toxicity Reanalysis Technical Memorandum, a more reasonable conclusion would be to set aside comparisons to the negative control treatments and rely upon comparisons to the reference envelope for this endpoint. In addition, it is inappropriate to remove this endpoint from consideration. Rather the strengths and weaknesses of this endpoint should be discussed in the uncertainty section of the BERA.

8) The draft BERA questions the reliability and utility of the Hyalella biomass results. EPA does not agree with these assertions. Site specific toxicity tests are considered surrogates for the actual benthic species exposed to site contaminants under EPA's national risk assessment guidance (OSWER 9285.7-28P). As the longest duration test of the toxicity tests performed at Portland Harbor, the Hyalella biomass test may be responding to contaminants that do not elicit toxicity during shorter exposures. The fact that the Hyalella biomass results are comparable to the other three empirical sediment toxicity tests performed at Portland Harbor can be demonstrated by a series of four maps prepared by EPA, and attached to these comments. The four maps show contours of Level 0, 1, 2 or 3 toxicity generated by a geographic information system nearest neighbor extrapolation approach. The areas of greatest toxicity are comparable for the four sets of toxicity tests, indicative that the tests are identifying toxicity in many of the same portions of the site. The Hyalella biomass results do show an elevated prevalence of Level 2 toxicity compared to the other three toxicity tests, which may be part of the basis for LWG's concerns. However, the elevated Level 2 toxicity prevalence is not a basis for questioning the reliability of the Hyalella biomass results. EPA's expanded reliability analyses and statistics indicate that the Hyalella biomass results are no less reliable than the other three toxicity tests, in contrast to the more limited reliability analyses performed by the LWG. Some of the observed variation in Hyalella biomass results may be due to the elevated chloride concentrations (or a compound for which chloride is a surrogate measure) in the laboratory control sediments relative to chloride content in Willamette River sediments. This appears related to the nutritional needs

of Hyalella in long term sediment exposures, and methods may be available to normalize results for this effect. EPA believes the time and effort LWG has spent questioning the reliability of the Hyalella biomass results would be better spent evaluating what the test results are actually telling risk assessors, which is comparable to the information provided by the other three toxicity test results.

Floating Percentile Model (FPM)

1) The description of the floating percentile model is insufficient to understand how the model was developed and implemented. For example, the discussion of the minimum detection threshold does not explain why many chemicals which met the minimum detection threshold of 30 were not included in the model. A detailed description of the model must be included in the revised BERA. In addition, during our January 22, 2010 meeting, a number of errors in the Benthic Toxicity Reanalysis Technical Memorandum were identified. Any such errors in the Benthic Toxicity Reanalysis Technical Memorandum should be corrected in the revised BERA.

2) The predictive models developed to support the BERA are based on incorrect interpretation of the sediment toxicity data. Furthermore, the data handling procedures for deriving sediment quality guidelines (SQGs) using the floating percentile model (FPM) are inappropriate. Non-detected results with detection limits (DLs) below conservative SQGs should have been used to develop and evaluate the models. These data were set aside according to the SQG-derivation methods described in the Technical Memorandum.

3) Chemicals Screened Out Prior to Model Development: These chemicals may be important toxicity drivers in localized areas of the site. There may not be enough information to include them in a site-wide predictive model. However, EPA requests plots of the hit and no-hit distribution for all chemicals that meet the minimum detection thresholds be presented in an appendix to the BERA.

4) The utility of the floating percentile model within the BERA has engendered more discussion among the government team than any other BERA review topic. A primary concern is that the FPM's very nature, which requires a subjective *a priori* definition of the false negative error rate and sensitivity before the model can be run places the FPM into the realm of risk management methodologies, not risk assessment methodologies. The proposed 20% false negative rate is higher than the combined prevalence of Level 2 plus Level 3 toxicity in the Chironomus survival and biomass, and the Hyalella survival sediment toxicity tests. This leads to the possibility that, depending on the distribution of chemical concentrations in sediment associated with toxic vs. nontoxic stations, the FPM yields sediment quality guidelines higher than the highest concentration eliciting toxicity, yet still meeting LWG's reliability goals for false negative rate and overall reliability of the FPM, an unacceptable SQG for EPA's use in the BERA. The discussions of overall model reliability given in the draft BERA and the benthic reanalysis memorandum do not address reliability of the individual sediment quality guidelines derived from the FPM.

5) EPA understands that the LWG followed the draft approach used in the past by the Regional Sediment Evaluation Team (RSET). However, concerns have been raised about using the

ANOVA test to limit constituents. The ANOVA test assumes that the data sets are normally distributed, which is unlikely the case for most chemicals. A defensible justification must be provided for any method to limit the number of chemicals included in the FPM analysis. Otherwise all chemicals should be included. In addition, the change in control normalization and biomass endpoint could significantly change the conclusions here for some chemicals.

6) Section 6.2.3, SQG Derivation: Calculating all error and reliability rates for each set of the initial SQGs by using the pooled endpoint (results shown in Table 6-9) will likely lead to inappropriately reduced reliability. The criteria for removing a chemical from the model dataset are inappropriate and likely led to a reduced set of SQGs as presented in Table 6-10. SQG values for chemicals removed after model development removes significant information in the weight of evidence framework and implies that there is no toxicity associated with the chemical, which is inaccurate.

7) Consistent with the approach outlined in the Approach for Assessing Risks to the Benthic Invertebrate Community at the Portland Harbor Superfund Site, the use of pooled endpoints should be avoided. Because sufficient data is available, individual endpoints should be used. Pooled endpoints may be presented in addition to the individual endpoints if the LWG so desires. This approach was documented in an EPA email dated October 14, 2008.

8) A sensitivity analysis that considers the order of the chemicals being run in the model should be performed and submitted. EPA understands that this analysis has already been performed.

9) To facilitate agency review of the chemical screening step, EPA requests the data files and statistical screening results for approval prior to proceeding with the future FPM modeling runs. Agreeing on the appropriate set of chemicals will avoid unnecessary development of SQGs using inappropriate datasets.

10) The sediment toxicity results and the FPM in the BERA should consider low medium and high thresholds. The analysis submitted to EPA only presents low and high. Ultimately, four bins of toxicity results should be considered – no toxicity and low, moderate and high.

11) The Floating Percentile Model (FPM) is not necessarily a peer reviewed model and requires subjective choices in how the model is run. Although this does not preclude the use of the FPM at the Portland Harbor site, the model should be run in the most objective manner possible. Unless the model can be reproduced and sufficient documentation of the model set up and implementation is provided, it is not suitable for use in the benthic risk evaluation. In general, the use of best professional judgment should be avoided to ensure that the model is reproducible. Examples of where BPF may have been inappropriately applied are summarized below:

- A minimum detection of 30 was established for chemicals to be used included in the model. However, this rule was sometimes violated. For example, in some cases chemicals detected more than 30 times were excluded and in other cases, chemicals detected less than 30 times were included. Justification for the cutoff of 30 detections should be provided. EPA cautions against deviating from any established frequency of detection threshold. However, if deviations are pursued, a logical approach should be applied rather than the use of professional judgment to ensure that the model is reproducible.

- The LRM establishes a false negative criteria of 20% false negatives, which seems high for use in a BERA, but not necessarily in the FS. The value of 20% must be justified. The order in which the chemicals are entered into the model should be presented and justified.

12) The draft BERA should clearly describe how chemicals such as dioxins/furans, DDx, chlordanes, and PAHs were summed for use in the FPM. It is unclear the effect on the model of including individual constituents (such as DDD, DDE and DDT) in addition to sums (total DDX), given the likely correlation between the individuals and the sum. The issue must be explicitly addressed. Also, "total toxic dioxin/furans" are included in the database without explanation. It is unclear if a TEQ approach was used, and if so, which TEFs were applied. For benthic toxicity, the appropriate sum should be total dioxins/furans.

13) It is unlikely that the appropriate statistical test was utilized to determine whether the chemical hit distribution was different from the no-hit distribution. If there is no difference between the hit and no-hit distributions for a given chemical, then it will not be possible to develop SQGs based on the FPM for that chemical. The ANOVA tests that were performed in the BERA were based on an assumption of normality. However, it is unlikely that the hit and no-hit distributions are in fact normally distributed. As a result, consistent with the ProUCL guidance, the statistical test should be specific to the distribution. The determination of whether a given chemical should be included in the FPM should be based on a comparison between the hit and no-hit distributions. An appropriate statistical test should be used to determine if there is a statistical difference between the hit and no-hit distributions. In most instances, this will require a non-parametric statistical test.

Logistic Regression Model (LRM)

1) The logistic regression model was inappropriately eliminated as a line of evidence in the draft BERA. EPA acknowledges some issues with the model. To address this concern, an updated logistic model has been developed by Jay Field. The updated model relies on site data only and uses organic carbon and grain size normalizations to improve model performance. Descriptions of the updated model, model results and supporting information are attached to these comments.

2) The reliability argument offered by LWG as their basis for eliminating logistic regression ignores the primary function of the LRM within the BERA. The Pmax statistic calculated by the LRM provides a non-threshold based measure of estimating the probability of sediment toxicity for a sample without empirical toxicity data. As such, the LRM provides the BERA with a true measure of risk, defined as the probability of adverse effect. Also, the LRM provides an indicator of the toxicity of chemical mixtures in sediment, as opposed to the chemical by chemical estimates of hazard provided by both the generic sediment quality benchmarks and the individual chemical benchmarks derived from the floating percentile model. Results from the LRM, including the Pmax estimates of risk to benthic invertebrates, must be fully presented and discussed in the BERA.

3) The updated model uses probability thresholds of 50% and 75% for individual chemicals in addition to Pmax values based on the LRM model for 30 chemicals. The probability thresholds for individual chemicals can be used to develop Portland Harbor specific sediment benchmarks for individual chemicals based on the LRM. See the attached updated LRM results.

Sediment Quality Guidelines

1) The SQG-evaluation procedures described in the Technical Memorandum are convoluted and are not focused on the key criteria of interest to risk managers for the site (i.e., false positive and false negative rates). As a result, many potentially useful predictive tools are labeled as unreliable and set aside. A more objective approach to SQG evaluation would likely have resulted in the identification of numerous SQGs that would be useful in the risk assessment and the risk management processes (e.g., < 20% incidence of toxicity to benthic invertebrates for low SQGs, considering survival or biomass of midge or amphipods; > 75% incidence of toxicity to benthic invertebrates for high SQGs, considering survival or growth of midge or amphipods).

2) The Technical Memorandum does not present the results of comparisons of whole-sediment chemistry to generic SQGs. In addition, methods used to make comparisons to site-specific SQGs are not transparent and are difficult to follow. This information is essential for identifying contaminants of concern in the study area and for determining which substances are causing or substantially contributing to benthic risk.

3) The procedures used to evaluate generic SQGs resulted in the conclusion that none of the high SQGs were reliable. It is likely that application of alternate reliability criteria would have resulted in different conclusions about the usability of the generic SQGs in the deterministic risk assessment for benthic invertebrates.

4) Section 6.2.4.1, Page 141, Comparison of Study Area Concentrations to Site-Specific SQGs: Comparison to SQGs should be presented here regardless of reliability in the BERA, with the reliability analyses of the SQGs given in the uncertainty analysis.

Reliability Analyses

1) The reliability analyses cannot be used to eliminate lines of evidence or individual sediment quality benchmarks from use in the BERA. Rather, the reliability analyses should be used in the uncertainty analysis and the weight of evidence evaluation. EPA’s comments build upon and augment the reliability analyses presented in the draft BERA and Benthic Toxicity Reanalysis Technical Memorandum and propose additional statistical evaluation tools to evaluate reliability of the sediment quality benchmarks. Additional information regarding reliability analyses is provided as an attachment.

2) The LWG’s application of reliability measures in the draft BERA is not a true measure of model or sediment quality benchmark reliability, because multiple reliability statistics (e.g. false negative rate and sensitivity) were subjectively set at predefined values in the floating percentile model to meet LWG-proposed risk management goals.

3) The draft BERA should use an objective approach to determine predictive model and sediment quality benchmark accuracy and reliability. Predictive models must be calibrated in such a manner that the sediment quality benchmarks derived from the models maximize the agreement between observed and model predicted toxicity for the 293 Portland Harbor stations for which co-occurring sediment toxicity and sediment chemistry data are available.

4) EPA's recommended solution to the effect of prevalence on reliability statistics is to base model reliability evaluations primarily on reliability statistics that can either be adjusted for prevalence, or whose values are not dependent on prevalence in the calibration dataset. Many of the statistics unaffected or unbiased by prevalence also have the useful property of assessing the extent to which models correctly predict toxicity at rates that are better than chance predictions of accuracy. To avoid information loss from not using all available information in a contingency table, reliability statistics are available that can be calculated using information from all contingency table cells. The reliability statistics used by LWG in the draft BERA, while providing useful information, do not make use of all available information in contingency tables. Statistics not utilized by the LWG in the draft BERA, but which are unaffected by prevalence or can be adjusted to account for prevalence effects, utilize all information in a contingency table, and/or which describe the improvement of model or benchmark predictions over the agreement between predicted and measured toxicity expected solely by chance include the odds ratio, Cohen's kappa, prevalence adjusted bias adjusted kappa (PABAK), the Hanssen-Kuipers discriminant, the normalized mutual information (NMI) statistic, and likelihood ratios. Of particular use in the BERA uncertainty analysis will be the reliability statistic called bias, which identifies whether a model or benchmark systematically over-or underestimates toxicity, as well as identifying the direction of the bias.

These statistics and the other statistics identified by EPA but not used in the draft BERA, should be used in addition to and in conjunction with the reliability statistics used by LWG to obtain a more complete and accurate picture of model and benchmark reliability. EPA is providing the LWG with a detailed reliability comment under separate cover, which includes the additional EPA recommended reliability statistics, their computation and use, and can provide the LWG with a spreadsheet that performs all the reliability calculations for a single sediment quality benchmark or predictive model if the counts of the number of true positive, true negative, false positive and false negative predictions are known.

5) The text states "none of the generic SQGs could reliably predict toxicity in Portland Harbor sediment, therefore the generic SQGs were not used in risk characterization." This depends on an agreed upon definition of "reliability" which still needs to be determined. The LWG reliability requirements are presented on page 138 (both false negatives and false positives should be below 20%, and overall reliability should be above 80%), and are based on the draft Washington State freshwater guidelines (Avocet, 2003). It is not a function of low SQGs (e.g. TECs) to predict toxicity accurately, instead, TECs are designed primarily to identify sediment where little or no toxicity is expected. This is also important since these criteria were used to evaluate the logistic regression, and is cited for the reason this line of evidence was not included (page 129).

6) The reliability analyses presented in the draft BERA provide useful information in the context of the BERA uncertainty analyses. However, this information should not be used to eliminate entire lines of evidence (e.g. logistic regression model) and many individual sediment quality benchmarks from the risk characterization portions of the BERA. As directed in the February 15, 2008 problem formulation, LWG must present the hazard quotients calculated for all sediment benchmarks directed for use, including generic benchmarks such as PECs and PELs, logistic regression derived and floating percentile model derived sediment quality benchmarks. The mean PEC quotient analyses will need to be recalculated based on the additional reliability statistics provided to LWG by EPA. It is likely that the reliability of the mean PEC quotient line of evidence can be improved by changing the threshold quotient from that given in the February 15, 2008 EPA problem formulation document.

7) Prevalence of toxicity, defined as the proportion of stations in the four sets of measured sediment toxicity data (Chironomus dilutus survival and biomass, Hyalella azteca survival and biomass) from 293 sampling locations in Portland Harbor that actually elicit toxicity, is low, ranging between 7% (Hyalella survival) to 25% (Hyalella biomass) of stations eliciting either Level 2 (moderate) or Level 3 (severe) toxicity, depending on which bioassay one is discussing. The low percentage of stations eliciting toxicity is an encouraging finding of the BERA, as it means between 75 – 93% of the 293 stations either elicit Level 0 (no toxicity), or Level 1 (low) levels of toxicity. Between 57% (Hyalella biomass) and 86% (Hyalella survival) of the 293 stations elicited no toxicity (Level 0).

The low prevalence of toxicity in the 293 stations with co-occurring sediment chemistry and sediment toxicity data used to develop the site specific floating percentile and logistic regression predictive toxicity models adversely affects the calculated values of the reliability statistics presented in the draft BERA and the benthic reanalysis memorandum, as well as their interpretation, and can also bias the reliability statistics. The prevalence effect is a statistical problem that directly results from the relatively low number of stations eliciting toxicity. The problem is not due to a lack of sampling data, nor is it a criticism of any particular predictive modeling approach or sediment quality benchmark.

8) All reliability statistics evaluated by the LWG in the draft BERA, as well as all additional reliability statistics recommended for use by EPA can be derived from a contingency table that tabulates the number of true positive, true negative, false positive and false negative predictions of toxicity made by any predictive model or any individual sediment chemical benchmark calibrated with or validated against the 293 stations with measured toxicity and sediment chemistry data. A broad suite of reliability measures including correct classification rate (number of non-toxic stations divided by number of predicted non-toxic and number of toxic stations divided by number of predicted toxic stations) should be used to support the reliability analysis and risk management decisions regarding the development of cleanup levels protective of the benthic community.

9) The overall floating percentile model reliability discussions in the draft BERA and Benthic Reanalysis Memorandum are useful discussions regarding uncertainties of the FPM in the BERA. Unfortunately, overall model reliability discussions do not address one of the two primary uncertainty and reliability question associated with the FPM, which is the reliability and

uncertainty associated with the individual chemical sediment quality guidelines derived from the FPM. Both the BERA and Benthic Reanalysis Tech Memo are silent on this topic. The absence of any analysis of the reliability and predictive accuracy of the individual sediment chemical guidelines derived with the FPM does not allow an answer to the question "what is the reliability of the individual chemical FPM derived sediment quality guidelines in predicting toxicity at stations without empirical toxicity data?". The reliability of and uncertainties associated with the individual chemical sediment quality guidelines derived from the floating percentile model must be presented and discussed in the next draft of the BERA, in addition to the overall model reliability.

10) Another primary concerns regarding uncertainties and reliability of sediment quality guidelines derived using the FPM regards the lack of conservatism of the FPM derived guidelines. Within the BERA, the emphasis is on identifying chemicals and locations that pose potentially unacceptable ecological risks to benthic biota. By setting the overall FPM acceptable false negative rate at a maximum of 20%, this implies that as many as one in five stations with sediment chemical concentrations will be incorrectly classified as nontoxic when in fact they are known to elicit toxicity based on the results of the empirical sediment toxicity tests used to develop the site specific FPM. This is an unacceptably high false negative error rate for the BERA. EPA believes that the allowable BERA false negative error rate should be lower than 20%, perhaps as low as 5%, even though this will result in an increase in the number of false positive errors. This is an acceptable approach within the BERA. If the floating percentile model is ultimately used to develop sediment remediation goals to be presented in the Record of Decision for the Portland Harbor site, it is up to the EPA risk manager to make the final determination regarding acceptable false negative error rates of remediation goals, which the risk manager may decide can be higher than the low false negative rate EPA proposes for use in the BERA.

11) Given the longstanding understanding and agreement between EPA and the LWG that the FPM is a line of evidence in the BERA, as stated in the February 15, 2008 problem formulation, the continued use of the FPM in the BERA is allowable. However, the reliability of the FPM must be evaluated thoroughly to determine the overall strength of the FPM relative to other lines of evidence. In addition, EPA requires reliability estimates from the FPM for individual chemical benchmarks, in addition to the previously presented overall model reliability estimate. EPA continues to have no objections to the use of the FPM in the feasibility study as one of several tools to assist risk managers in identifying remediation goals for sediment.

Weight of Evidence

1) The weight-of-evidence (WOE) approach presented in the BERA is inconsistent with the WOE approach presented in the Problem Formulation. Page 59 of the EPA Problem Formulation states that the weight of evidence evaluation:

> “(C)an be used by itself to help identify which LOEs have the most scientific reliability and, hence, are of most use in applying risk assessment results to making remedial decisions. WOE approaches can also be used to integrate this information along with the *magnitude of response* of each measurement for each chemical stressor and the

> *concurrence among measurement endpoints*. For the Portland Harbor BERA, however, EPA developed a WOE framework that, for the time being, only focuses on evaluating the scientific reliability or "weighting" of each measurement endpoint or LOE. *Our ultimate goal was to develop a method to help identify and rank which LOEs for each receptor provide the most scientifically reliable indication of the status of each assessment endpoint from exposure to COPCs at the site and, hence, which might be most useful for making risk management decisions.*

The EPA Problem Formulation further states on Page 61:

> "(T)he WOE Framework should be applied to the results of the risk characterization to provide an overall assessment of the strength of each measurement endpoint. Risk characterization results should be summarized in a table and each line of evidence should be identified as a high, medium or low weight. For receptors with multiple measurement endpoints, this information may useful from the standpoint of assessing uncertainty and making effective risk management decisions. Because this WOE Framework is focused on evaluating and documenting the general scientific reliability of each LOE, it can be used to help resolve potential inconsistencies in risk estimates among multiple LOEs for the same receptor. Particularly in cases where risk estimates among multiple LOEs directly contradict one another, the WOE Framework can help identify which risk estimates are the most scientifically reliable and, hence, most likely to be a correct and meaningful assessment of risk for that receptor or Assessment Endpoint."

Based on the description of the WOE framework in the problem formulation, it is clear that all lines of evidence identified for evaluation in the benthic risk evaluation should be used to assess risk to the benthic community. The WOE evaluation should then present information that can be used by the risk managers to determine which lines of evidence should be used to establish cleanup goals for the Portland Harbor site

2) The BERA did not follow the guidance provided by EPA in the Problem Formulation for the use of sediment quality guidelines. EPA suggested use of higher SQG values and did not suggest comparisons to empirical toxicity. The BERA also appears to use a different approach in applying the WOE framework (Section 6.7, page 213). EPA's minimum requirement is that evaluation of SQGs and the WOE evaluation should be consistent with the EPA Problem Formulation. We have no objection to LWG presenting information and analyses above the minimum EPA requirement, although we reserve the right to disagree with and ultimately not consider any such additional analyses in either the BERA or the FS.

Bivalve Growth and Mortality

Section 6.1.2.1, Page 137, Toxicity Assessment Based on Bivalve Growth and Mortality: This evaluation is new, and the raw data should be submitted (it was not submitted in previous documents). These data should be included as a line of evidence in the final risk characterization.

Individual Endpoint supplement

The April 2, 2010 Site Specific SQGs based on Individual Bioassay Endpoints supplement presents the results of the individual endpoints. However, the number of chemicals evaluated based on individual endpoints is fewer than the chemicals evaluated based on the pooled endpoint. A stated above, the results should be evaluated on an individual chemical basis. Further elaboration of this approach beyond what is presented in the April 2, 2010 document should be provided.

Editorial

Global editorial change: the term Hyalella growth and similar terms should be changed to Hyalella biomass when referring to the sediment toxicity tests performed as part of the Portland Harbor BERA.

The discussion in the first paragraph of Section 6.2.1 correctly explains the rationale for using 293 sediment toxicity test station results in developing predictive models, 269 of which are in the study area, and 13 of which are from locations subsequently dredged, and thus not part of the empirical sediment toxicity analyses. Given the difference in sample numbers used in various parts of the BERA (i.e. 293 stations for predictive toxicity model development, 256 for empirical toxicity discussions), it would be useful to highlight this discussion in a green call out box, so this discussion can be more readily found by the reader.